An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls

Yinchu Zhu

(joint with Victor Chernozhukov and Kaspar Wüthrich)

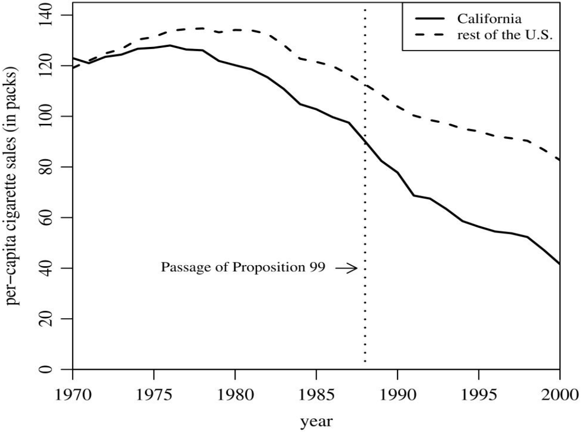
June 1, 2018

This Paper

- New and theoretically justified inference methods for counterfactual and synthetic control (CSC) methods
  - Synthetic controls
  - Penalized regression models
  - Factor models
  - Times series models
  - …
- Key feature: double justification
  - Exact finite sample validity under strong assumptions
  - Approximate validity under weak assumptions

Classical Synthetic Control

- What is the causal effect of California's Proposition 99?
  - Proposition 99: Anti-tabacco legislation which increased cigarette excise tax, earmarked tax revenues to health and anti-smoking education, funded anti-smoking campaigns, spurred clean indoor-air ordinances



Classical Synthetic Control

- Classic approach: differences-in-differences.
  - But what state should we use as control for California?
- Synthetic control (SC) approach (Abadie, Diamond, and Hainmueller, 2010; Doudchenko and Imbens, 2016; Athey, Bayati, Doudchenko, Imbens, and Khosravi, 2017): construct a synthetic California.
  - Synthetic California = weighted combination of other states
  - Weights are constructed to maximize pre-treatment fit.

Setup and Notation

Aggregate panel data setup:

- $J+1$ units (US states), unit $j=1$ (California) is treated, units $j=2,\ldots,J+1$ are the control units, which constitute the donor pool.
- T_0 pre-treatment periods, T_* post-treatment periods.
- Y_{jt}: observed outcome, Y_{jt}^N: counterfactual outcome without treatment.
- $Y_{jt} = Y_{jt}^N + \alpha_t D_{jt}$ with

$$D_{jt} = \begin{cases} 1 & j=1 \text{ and } t > T_0 \\ 0 & \text{otherwise} \end{cases}$$

- We will omit covariates for simplicity throughout this talk.

Estimating Synthetic California

- Assume (Doudchenko and Imbens (2016)) that

$$Y_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt}^N \qquad \text{for } t = 1, ..., T,$$

where $T = T_0 + T_*$ and the weights $w = (w_2, \ldots, w_{J+1})'$ can be estimated by

$$\hat{w} = \arg\min_w \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \quad \text{s.t.} \;\; 0 \leq w_j \leq 1, \; \sum_{j=2}^{J+1} w_j = 1$$

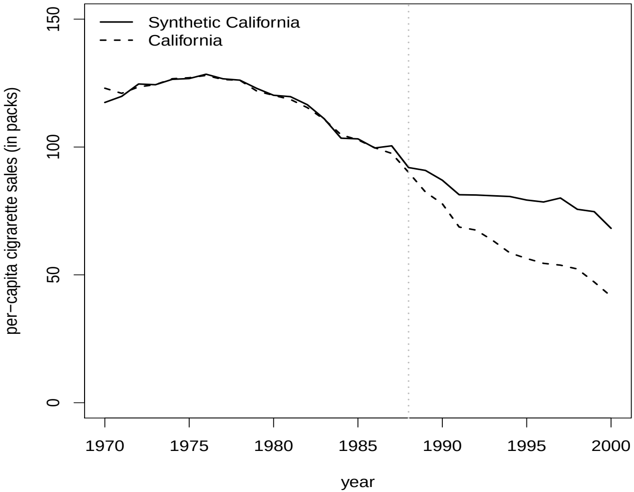
- The treatment effect α_t can be estimated as for $t \geq T_0 + 1$,

$$\hat{\alpha}_t = Y_{1t} - \hat{Y}_{1t}^N = Y_{1t} - \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$$

Estimated Weights: Synthetic California

AL	0	NV	0.20
AR	0	NH	0.05
CO	0.01	NM	0
CT	0.11	NC	0
DE	0	ND	0
GA	0	OH	0
ID	0	OK	0
IL	0	PA	0
IN	0	RI	0
IA	0	SC	0
KS	0	SD	0
KY	0	TN	0
LA	0	TX	0
ME	0	UT	0.39
MN	0	VT	0
MS	0	VA	0
MO	0	WV	0
MT	0.23	WI	0
NE	0	WY	0

Results: Synthetic California



Proposed Inference Procedure: Simple Case

- Assume (Doudchenko and Imbens (2016)) that

$$\left.\begin{aligned} Y_{1t}^N &= \sum_{j=2}^{J+1} w_j Y_{jt} + u_t \\ Y_{1t}^I &= \sum_{j=2}^{J+1} w_j Y_{jt} + \alpha_t + u_t \end{aligned}\right| \quad E(u_t) = 0, \quad t = 1, \ldots, T$$

where $\{u_t\}$ is i.i.d residual and α_t is the treatment effect at t.

- Suppose that $T_* = 1$. We want to test $H_0 : \ \alpha_T = \alpha_T^o$
- Under H_0, $\{Y_{1t}^N\}_{t=1}^T$ is observed because

$$Y_{1T}^N = Y_{1T}^I - \alpha_T^o = Y_{1T} - \alpha_T^o$$

- Suppose further that w is known.
- Idea: compute $u = (u_1, \ldots, u_T)$ with $u_t = Y_{1t}^N - \sum_j w_j Y_{jt}$ and compare u_T to $(u_1, \ldots, u_{T_0})$.
- Related to Andrews (2003).

Proposed Inference Procedure: Simple Case

- Define the test statistic $S(u) := |u_T|$ for $u = (u_1, ..., u_T)$.
- Let Π denote the set of all permutations of $\{1, \ldots, T\}$ and let $u_\pi = (u_{\pi(1)}, \ldots, u_{\pi(T)})$ denote the vector of errors permuted by $\pi \in \Pi$.
- Let $n = |\Pi|$. A permutation p-value can be obtained as

$$\hat{p} = \frac{1}{n} \sum_{\pi \in \Pi} \mathbf{1} \left\{ S\left(u_\pi\right) \geq S(u) \right\} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1} \left\{ |u_t| \geq |u_T| \right\}.$$

- This procedure achieves exact finite sample validity: for $\alpha \in (0, 1)$,

$$P(\hat{p} \leq \alpha) \leq \alpha$$

- Related to randomization/permutation tests (e.g., Romano (1990), Lehmann and Romano (2005)) and conformal prediction (e.g., Vovk, Gammerman, and Shafer (2005), Vovk, Nouretdinov, and Gammerman (2009), Lei, G'Sell, Rinaldo, Tibshirani, and Wasserman (2017)).

Exact Validity: Proof

- Let $\{S^{(j)}(u)\}_{j=1}^{n}$ denote the ordered values of $\{S(u_\pi) : \pi \in \Pi\}$.
- Observe that $\mathbf{1}(\hat{p} \leq \alpha) = \mathbf{1}(S(u) > S^{(k)}(u))$ where $k = \lceil n(1-\alpha) \rceil$.
- Because $S^{(k)}(u) = S^{(k)}(u_\pi)$ for all $\pi \in \Pi$,

$$\sum_{\pi \in \Pi} \mathbf{1}(S(u_\pi) > S^{(k)}(u_\pi)) = \sum_{\pi \in \Pi} \mathbf{1}(S(u_\pi) > S^{(k)}(u)) \leq \alpha n.$$

- The i.i.d assumption means that u and u_π have the same distribution $\forall \pi \in \Pi$.
- Hence, $\mathbf{1}(S(u) > S^{(k)}(u))$ is equal in law to $\mathbf{1}(S(u_\pi) > S^{(k)}(u_\pi))$ and we have

$$\begin{aligned}
\alpha &\geq \frac{1}{n} E\left(\sum_{\pi \in \Pi} \mathbf{1}(S(u_\pi) > S^{(k)}(u_\pi)) \right) \\
&= E\left(\mathbf{1}(S(u) > S^{(k)}(u)) \right) = E\left(\mathbf{1}(\hat{p} \leq \alpha) \right).
\end{aligned}$$

Preview of Main Results

- Sharp null hypothesis:

$$H_0 : (\alpha_{T_0+1}, \ldots, \alpha_T) = (\alpha^o_{T_0+1}, \ldots, \alpha^o_T)$$

- General class of counterfactual models:

$$Y^N_{1t} = P^N_t + u_t$$

  Formulation nests synthetic controls, factor models, penalized regression models, times series models, ...
- Propose an inference procedure based on estimated P^N_t and u_t.
- Inference procedure has a double justification:
  - Exact validity under exchangeability of $\{\hat{u}_t\}$
  - Approximate validity under weak assumptions on $\hat{P}^N_t$

Literature: Inference for Synthetic Control Models

- Finite population approaches
  - Randomization/permutation inference approaches for testing sharp null hypotheses
  - Key assumptions: treatment assignment (or timing of intervention) is random, potential outcomes are fixed but unknown
  - Some references: Abadie and Gardeazabal (2003), Abadie, Diamond, and Hainmueller (2010), Abadie, Diamond, and Hainmueller (2015), Doudchenko and Imbens (2016), Hahn and Shi (2016), Firpo and Possebom (2017)

- Asymptotic approaches
  - Asymptotic inference for average effects based on estimated counterfactuals
  - Key assumptions: $T_0 \to \infty$, $T_1 \to \infty$, $J \to \infty$ plus (sparsity or factor structure)
  - Some references: Hsiao, Steve Ching, and Ki Wan (2012), Gobillon and Magnac (2016), Chan and Kwok (2016), Carvalho, Masini, and Medeiros (2017), Athey, Bayati, Doudchenko, Imbens, and Khosravi (2017), Li (2017, 2018)

Setup

Assumption

There exists a sequence of mean unbiased proxies $\{P_t^N\}$ *such that*

$$\left.\begin{array}{l} Y_{1t}^N = P_t^N + u_t \\ Y_{1t}^I = Y_{1t}^N + \alpha_t \end{array}\right| \quad E(u_t) = 0, \quad t = 1, \ldots, T$$

where $T = T_0 + T_*$.

Implicit assumptions:

- Availability of an estimator $\hat{P}_t^N$
- Stochastic shock sequence is stationary under the intervention.

Example I: Constrained LASSO

- Model (Doudchenko and Imbens (2016)):

$$P_t^N = \mu + \sum_{j=2}^{J+1} w_j Y_{jt}, \text{ where } \|w\|_1 \leq 1$$

- Estimator:

$$\hat{P}_t^N = \hat{\mu} + \sum_{j=2}^{J+1} \hat{w}_j Y_{jt},$$

where $(\hat{\mu}, \hat{w})$ are obtained using constrained Lasso

$$(\hat{\mu}, \hat{w}) = \arg\min_{(\mu, w)} \sum_{t=1}^{T_0} \left(Y_{1t} - \mu - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \text{ s.t. } \|w\|_1 \leq 1$$

Example II: Factor Models

- Assume a factor model for all units: for $j = 1, \ldots, T$,

$$Y_{jt}^N = \lambda_j' F_t + u_{jt}$$

- Model:

$$P_t^N = \lambda_1' F_t$$

- Estimator:

$$\hat{P}_t^N = \hat{\lambda}_1' \hat{F}_t,$$

where $\hat{\lambda}_1$ and $\hat{F}_t$ are obtained using standard PCA.

Example III: Penalized Regression

- Model:

$$P_t^N = \mu + \sum_{j=2}^{J+1} w_j Y_{jt}.$$

- Estimator:

$$\hat{P}_t^N = \hat{\mu} + \sum_{j=2}^{J+1} \hat{w}_j Y_{jt},$$

where

$$(\hat{\mu}, \hat{w}) = \arg\min_{(\mu,w)} \sum_{t=1}^{T_0} \left(Y_{1t} - \mu - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 + \mathcal{P}(w)$$

Further Examples

- Interactive FE models
- Matrix completion models
- Dynamic models such as AR-models, neural net, fused times-series-panel models
- …

Hypothesis of Interest and Test Statistic

- Hypothesis of interest:

$$H_0 : (\alpha_{T_0+1}, \ldots, \alpha_T) = \left(\alpha^o_{T_0+1}, \ldots, \alpha^o_T\right)$$

- Data under the null: $Z = (Z_1, \ldots, Z_T)'$, where $Z_t = \left(Y^N_{1t}, Y^N_{2t}, \ldots, Y^N_{J+1t}\right)'$ and

$$Y^N_{1t} = \begin{cases} Y^N_{1t} & t \leq T_0 \\ Y_{1t} - \alpha^o_t & t > T_0. \end{cases}$$

- Obtain $\hat{P}^N_t$ under the null (i.e., using Z) and compute

$$\hat{u} = (\hat{u}_1, \ldots, \hat{u}_T)', \quad \hat{u}_t = Y^N_{1t} - \hat{P}^N_t, \quad t = 1, \ldots, T.$$

- Test statistic (other choices are possible)

$$S(\hat{u}) = \left(\frac{1}{\sqrt{T_*}} \sum_{t=T_0+1}^{T} |\hat{u}_t|^q \right)^{1/q}.$$

In applications, we set $q = 1$.

Computing p-Values by Permuting Residuals

- For each $\pi \in \Pi$, let $\hat{u}_\pi = (\hat{u}_{\pi(1)}, \ldots, \hat{u}_{\pi(T)})'$ denote the vector of permuted residuals.
- The p-value is

$$\hat{p} = 1 - \hat{F}(S(\hat{u})), \quad \text{where } \hat{F}(x) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbf{1}\{S(\hat{u}_\pi) < x\}.$$

- Can test $H_0 : \alpha_t = \alpha_t^o$ for $t > T_0$ using $Z = (Z_1, \ldots, Z_{T_0}, Z_t)'$
- Confidence sets can be constructed by test inversion.

Permutations

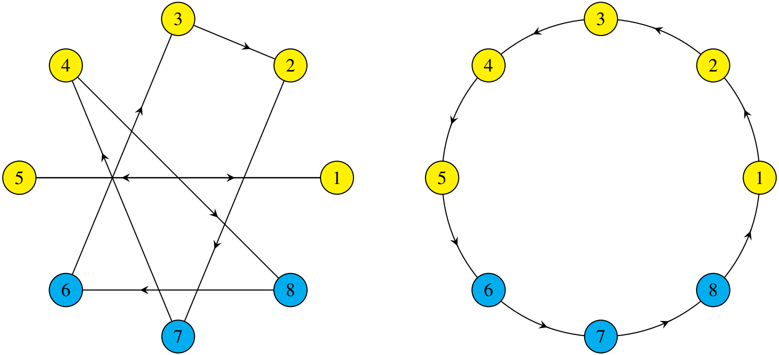
We consider two types of permutations:

- The set of all permutations, which we call *i.i.d. permutations*, Π_{all}.
  $\rightarrow$ more elements, requires i.i.d. $\{u_t\}$
- The set of all (overlapping) *moving block permutations*, $\Pi_{\rightarrow}$. The elements of $\Pi_{\rightarrow}$ are indexed by j and the permutation π_j does the following:

$$\pi_j(i) = \begin{cases} i+j & \text{if } i+j \leq T \\ i+j-T & \text{otherwise.} \end{cases}$$

  $\rightarrow$ fewer elements, allows weakly dependent $\{u_t\}$

Graphical Illustration Π_{all} and $\Pi_{\rightarrow}$



- Example of $\Pi_{\rightarrow}$. Consider $T = 5$, i.e., $\{1, 2, 3, 4, 5\}$.
    - $\{1, 2, 3, 4, 5\} \rightarrow \{2, 3, 4, 5, 1\}$
    - $\{1, 2, 3, 4, 5\} \rightarrow \{3, 4, 5, 1, 2\}$
    - $\{1, 2, 3, 4, 5\} \rightarrow \{4, 5, 1, 2, 3\}$
    - $\{1, 2, 3, 4, 5\} \rightarrow \{5, 1, 2, 3, 4\}$

Exact Validity under Strong Assumptions

Theorem (Exact Validity)

Suppose that the null hypothesis is true. Suppose that $\{\hat{u}_t\}_{t=1}^{T}$ *is exchangeable with respect to* Π *under the null hypothesis. Then*

$$P\left(\hat{p} \leq \alpha\right) \leq \alpha.$$

When can we expect exchangeable $\hat{u}$?

Lemma

Suppose that $\hat{u}_t = g(Z_t, \hat{\beta})$, *where* $\hat{\beta} = \hat{\beta}(\{Z_t\}_{t=1}^T)$ *is an estimator for the model parameter. If*

- $\{Z_t\}_{t=1}^T$ *is i.i.d or exchangeable under* Π.
- $\hat{\beta}(\{Z_t\}_{t=1}^T)$ *is invariant under permutations.*

then $\hat{u} = (\hat{u}_1, ..., \hat{u}_T)$ *is exchangeable under* Π.

When can we expect exchangeable $\hat{u}$?

Lemma

Suppose that $\hat{u}_t = g(Z_t, \hat{\beta})$, *where* $\hat{\beta} = \hat{\beta}(\{Z_t\}_{t=1}^T)$ *is an estimator for the model parameter. If*

- $\{Z_t\}_{t=1}^T$ *is i.i.d or exchangeable under* Π.
- $\hat{\beta}(\{Z_t\}_{t=1}^T)$ *is invariant under permutations.*

then $\hat{u} = (\hat{u}_1, ..., \hat{u}_T)$ *is exchangeable under* Π.

- Most of estimates will be invariant to permutations of the data, e.g., LASSO, Synthetic Control, etc.

When can we expect exchangeable $\hat{u}$?

Lemma

Suppose that $\hat{u}_t = g(Z_t, \hat{\beta})$, *where* $\hat{\beta} = \hat{\beta}(\{Z_t\}_{t=1}^T)$ *is an estimator for the model parameter. If*

- $\{Z_t\}_{t=1}^T$ *is i.i.d or exchangeable under* Π.
- $\hat{\beta}(\{Z_t\}_{t=1}^T)$ *is invariant under permutations.*

then $\hat{u} = (\hat{u}_1, ..., \hat{u}_T)$ *is exchangeable under* Π.

- Most of estimates will be invariant to permutations of the data, e.g., LASSO, Synthetic Control, etc.
- Model-free performance guarantee: size control **even if estimator is misspecified or inconsistent**.

Approximate Validity: High-level Conditions

Assumption (Weak Dependence)

Assume

1. $\{u_t\}_{t=1}^T$ *are i.i.d., or*
2. $\{u_t\}_{t=1}^T$ *are stationary and strong mixing.*

Assumption (Consistency)

Let there be sequences of constants δ_T *and* γ_T *converging to zero. Assume that with probability* $1-\gamma_T$,

(1) *the mean squared estimation error is small,* $\|\hat{P}^N - P^N\|_2^2/T \leq \delta_T^2$;

(2) *for* $T_0 + 1 \leq t \leq T$, *the pointwise errors are small,* $|\hat{P}_t^N - P_t^N| \leq \delta_T$;

Approximate Validity under Weak Assumptions

Theorem (**Approximate Validity**)

We assume that T_* is fixed and $T_0 \to \infty$. Suppose that Assumption 5 holds. Impose Assumption 4.1 if Π_{all} is used. Impose Assumption 4.2 if $\Pi_{\to}$ is used. Then,

$$|P(\hat{p} \leq \alpha) - \alpha| \leq C\left((T_*/T_0)^{1/4}\log T + \delta_T + \sqrt{\delta_T} + \gamma_T\right),$$

where $C > 0$ is a constant (not depending on T_0).

Approximate Validity under Weak Assumptions

Theorem (**Approximate Validity**)

We assume that T_* is fixed and $T_0 \to \infty$. Suppose that Assumption 5 holds. Impose Assumption 4.1 if Π_{all} is used. Impose Assumption 4.2 if $\Pi_{\to}$ is used. Then,

$$|P(\hat{p} \leq \alpha) - \alpha| \leq C\left((T_*/T_0)^{1/4} \log T + \delta_T + \sqrt{\delta_T} + \gamma_T\right),$$

where $C > 0$ is a constant (not depending on T_0).

- Clearly, $|P(\hat{p} \leq \alpha) - \alpha| \to 0$ as $T_0 \to \infty$.

Approximate Validity under Weak Assumptions

Theorem (**Approximate Validity**)

We assume that T_* is fixed and $T_0 \to \infty$. Suppose that Assumption 5 holds. Impose Assumption 4.1 if Π_{all} is used. Impose Assumption 4.2 if $\Pi_{\to}$ is used. Then,

$$|P(\hat{p} \leq \alpha) - \alpha| \leq C\left((T_*/T_0)^{1/4} \log T + \delta_T + \sqrt{\delta_T} + \gamma_T\right),$$

where $C > 0$ is a constant (not depending on T_0).

- Clearly, $|P(\hat{p} \leq \alpha) - \alpha| \to 0$ as $T_0 \to \infty$.
- The convergence is uniform in underlying distribution since the above bound holds in finite samples.

Proof Sketch

- The following two conditions imply the results. Let $n \to \infty$.

(E) With probability $1 - \gamma_{1n}$:

$$\tilde{F}(x) := \frac{1}{n} \sum_{\pi \in \Pi} \mathbf{1}\{S(u_\pi) < x\},$$

is *approximately ergodic* for $F(x) = P\left(S(u) < x\right)$, namely

$$\sup_{x \in \mathbb{R}} \left| \tilde{F}(x) - F\left(x\right) \right| \leq \delta_{1n},$$

(A) With probability $1 - \gamma_{2n}$:

(1) $n^{-1} \sum_{\pi \in \Pi} \left[S(\hat{u}_\pi) - S(u_\pi)\right]^2 \leq \delta_{2n}^2$;

(2) $|S(\hat{u}) - S(u)| \leq \delta_{2n}$;

- Show that (E) holds for Π_{all} with i.i.d. $\{u_t\}$ and for $\Pi_{\to}$ with weakly dependent $\{u_t\}$.
- Show that assumption (A) is implied by the assumptions on $\hat{P}_t^N$

Sufficient Conditions (constrained LASSO)

Sufficient Conditions (constrained LASSO)

- Fix the parameter space $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$.

Sufficient Conditions (constrained LASSO)

- Fix the parameter space $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$.
  - Example: synthetic control uses $\mathcal{W} = \{v \in \mathbb{R}^J : v \geq 0 \text{ and } \|v\|_1 = 1\}$.

Sufficient Conditions (constrained LASSO)

- Fix the parameter space $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$.
  - Example: synthetic control uses $\mathcal{W} = \{v \in \mathbb{R}^J : v \geq 0 \text{ and } \|v\|_1 = 1\}$.
- Assume that $P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}$ for some $w \in \mathcal{W}$.

Sufficient Conditions (constrained LASSO)

- Fix the parameter space $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \ \|v\|_1 \leq K\}$.
  - Example: synthetic control uses $\mathcal{W} = \{v \in \mathbb{R}^J : \ v \geq 0 \text{ and } \|v\|_1 = 1\}$.
- Assume that $P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}$ for some $w \in \mathcal{W}$.
- Estimator: $\hat{P}_t^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$, where

$$\hat{w} = \arg\min_{w} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \quad \text{s.t.} \ \ w \in \mathcal{W}$$

Sufficient Conditions (constrained LASSO)

- Fix the parameter space $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$.
  - Example: synthetic control uses $\mathcal{W} = \{v \in \mathbb{R}^J : v \geq 0 \text{ and } \|v\|_1 = 1\}$.
- Assume that $P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}$ for some $w \in \mathcal{W}$.
- Estimator: $\hat{P}_t^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$, where

$$\hat{w} = \arg\min_{w} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \quad \text{s.t. } w \in \mathcal{W}$$

- **Lemma:** Assumption (Consistency) is satisfied if

Sufficient Conditions (constrained LASSO)

- Fix the parameter space $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$.
  - Example: synthetic control uses $\mathcal{W} = \{v \in \mathbb{R}^J : v \geq 0 \text{ and } \|v\|_1 = 1\}$.
- Assume that $P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}$ for some $w \in \mathcal{W}$.
- Estimator: $\hat{P}_t^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$, where

$$\hat{w} = \arg\min_{w} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \quad \text{s.t. } w \in \mathcal{W}$$

- **Lemma:** Assumption (Consistency) is satisfied if
  - Moment conditions and weak dependence on u_t and Y_{jt}.

Sufficient Conditions (constrained LASSO)

- Fix the parameter space $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$.
  - Example: synthetic control uses $\mathcal{W} = \{v \in \mathbb{R}^J : v \geq 0 \text{ and } \|v\|_1 = 1\}$.
- Assume that $P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}$ for some $w \in \mathcal{W}$.
- Estimator: $\hat{P}_t^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$, where

$$\hat{w} = \arg\min_w \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \quad \text{s.t. } w \in \mathcal{W}$$

- **Lemma:** Assumption (Consistency) is satisfied if
  - Moment conditions and weak dependence on u_t and Y_{jt}.
  - $\log J = o(T^c)$, where $c > 0$ is a constant. (Allow $J \gg T$.)

Sufficient Conditions (constrained LASSO)

- Fix the parameter space $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$.
  - Example: synthetic control uses $\mathcal{W} = \{v \in \mathbb{R}^J : v \geq 0 \text{ and } \|v\|_1 = 1\}$.
- Assume that $P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}$ for some $w \in \mathcal{W}$.
- Estimator: $\hat{P}_t^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$, where

$$\hat{w} = \arg\min_{w} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \quad \text{s.t. } w \in \mathcal{W}$$

- **Lemma:** Assumption (Consistency) is satisfied if
  - Moment conditions and weak dependence on u_t and Y_{jt}.
  - $\log J = o(T^c)$, where $c > 0$ is a constant. (Allow $J \gg T$.)
- No sparsity requirement on w.

Sufficient Conditions (constrained LASSO)

- Fix the parameter space $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$.
  - Example: synthetic control uses $\mathcal{W} = \{v \in \mathbb{R}^J : v \geq 0 \text{ and } \|v\|_1 = 1\}$.
- Assume that $P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}$ for some $w \in \mathcal{W}$.
- Estimator: $\hat{P}_t^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$, where

$$\hat{w} = \arg\min_w \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2 \quad \text{s.t. } w \in \mathcal{W}$$

- **Lemma:** Assumption (Consistency) is satisfied if
  - Moment conditions and weak dependence on u_t and Y_{jt}.
  - $\log J = o(T^c)$, where $c > 0$ is a constant. (Allow $J \gg T$.)
- No sparsity requirement on w.
- Once we choose $\mathcal{W}$, we do not need to choose tuning parameters (unlike Lasso).

Sufficient Conditions (in Paper)

The paper also provides sufficient conditions for the following models:

- Pure factor models
- Interactive FE models
- Matrix completion via nuclear norm penalization
- (Non-)linear AR models
- Fused models: panel data models with AR errors

Competitors

- Classical SC model w/o intercept (a special case of Constrained LASSO)
- Simple factor model w/o covariates

Simulation Setup

- DGPs similar to Hahn and Shi (2016).
  - Factor model for controls + SC model for treated unit
    - (a) Sparse weights
    - (b) Dense weights
  - Factor model for all units
    - (a) Common support
    - (b) No common support
- Focus on: $H_0 : \alpha_{T_0+1} = 0$

DGP1a,b: SC Models

- For $j = 2, ..., J + 1$, let

$$Y_{jt} = \alpha_j + \theta_t + \gamma_j' \delta_t + \epsilon_{jt},$$

where $\alpha_j = j/J$, $\theta_t \sim N(0, 1)$, $\delta_t \sim N(0, 1)$ is a scalar, $\gamma_j = j/J$.

- The treated outcome is generated as

$$Y_{1t} = \begin{cases} \sum_{j=2}^{J+1} w_j Y_{jt} + u_t & \text{if } t = 1, ..., T_0 \\ \alpha_t + \sum_{j=2}^{J+1} w_j Y_{jt} + u_t & \text{if } t = T_0 + 1, \ldots, T \end{cases}$$

where w is sparse (DGP1a) or dense (DGP1b).

- Two different variants:
  - $u_t \overset{iid}{\sim} N(0, 1)$ and $\epsilon_{jt} \overset{iid}{\sim} N(0, 1)$
  - AR(1) models for u_t and ϵ_{jt}

DGP2a,b: Factor Models

- For $j = 2, ..., J + 1$, let

$$Y_{jt} = \alpha_j + \theta_t + \gamma_j' \delta_t + \epsilon_{jt},$$

where $\alpha_j = j/J$, $\theta_t \sim N(0,1)$, $\delta_t \sim N(0,1)$ is a scalar, $\gamma_j = j/J$.

- The treated outcome is generated as (DGP2a)

$$Y_{1t} = 0.5 + \theta_t + 0.5\delta_t + \epsilon_{1t},$$

or as (DGP2b)

$$Y_{1t} = -0.5 + \theta_t - 0.5\delta_t + \epsilon_{1t},$$

- Two different variants:
  - $\epsilon_{jt} \overset{iid}{\sim} N(0,1)$
  - AR(1) model for ϵ_{jt}

Size DGP1a (Sparse SC Model)

	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.09	0.09	0.09	0.09	0.09	0.08	0.09	0.09	0.09
$T_0 = 50$	0.10	0.10	0.09	0.10	0.09	0.10	0.09	0.10	0.10
$T_0 = 100$	0.10	0.11	0.10	0.10	0.10	0.10	0.10	0.11	0.10
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.10	0.11	0.11	0.12	0.12	0.12	0.11	0.11	0.11
$T_0 = 50$	0.12	0.12	0.12	0.13	0.12	0.12	0.12	0.13	0.12
$T_0 = 100$	0.12	0.11	0.11	0.13	0.11	0.10	0.13	0.11	0.11

Size DGP1b (Dense SC Model)

	Synthetic control			Factor model			Constrained Lasso		
i.i.d. data with $\rho_\epsilon = \rho_u = 0$									
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.10	0.10	0.10	0.10	0.09	0.10	0.10	0.09	0.10
$T_0 = 50$	0.09	0.10	0.08	0.09	0.09	0.09	0.09	0.11	0.08
$T_0 = 100$	0.11	0.09	0.11	0.11	0.09	0.10	0.11	0.09	0.10
Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$									
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.11	0.12	0.10	0.13	0.13	0.14	0.12	0.13	0.10
$T_0 = 50$	0.12	0.13	0.12	0.11	0.12	0.12	0.13	0.12	0.12
$T_0 = 100$	0.11	0.10	0.11	0.10	0.11	0.11	0.11	0.11	0.11

Size DGP2a (Factor Model w/ Common Support)

	Synthetic control			Factor model			Constrained Lasso		
i.i.d. data with $\rho_\epsilon = \rho_u = 0$									
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.09	0.09	0.10	0.09	0.09	0.10	0.10	0.09	0.10
$T_0 = 50$	0.11	0.09	0.10	0.11	0.10	0.10	0.11	0.09	0.10
$T_0 = 100$	0.09	0.11	0.10	0.09	0.10	0.10	0.09	0.11	0.11
Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$									
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.12	0.12	0.11	0.12	0.13	0.13	0.12	0.12	0.11
$T_0 = 50$	0.13	0.12	0.12	0.13	0.12	0.12	0.13	0.14	0.13
$T_0 = 100$	0.12	0.12	0.12	0.12	0.11	0.11	0.13	0.12	0.12

Size DGP2b (Factor Model w/o Common Support)

	Synthetic control			Factor model			Constrained Lasso		
i.i.d. data with $\rho_\epsilon = \rho_u = 0$									
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.10	0.11	0.09	0.09	0.10	0.09	0.10	0.10	0.09
$T_0 = 50$	0.10	0.11	0.10	0.10	0.11	0.09	0.10	0.11	0.09
$T_0 = 100$	0.11	0.09	0.11	0.10	0.10	0.12	0.11	0.10	0.10
Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$									
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.11	0.12	0.11	0.12	0.12	0.11	0.11	0.10	0.09
$T_0 = 50$	0.10	0.12	0.11	0.10	0.11	0.11	0.11	0.12	0.11
$T_0 = 100$	0.10	0.11	0.13	0.11	0.10	0.11	0.11	0.12	0.10

Power DGP1a (Sparse SC Model)

i.i.d. data with $\rho_\epsilon = \rho_u = 0$

	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.49	0.45	0.43	0.38	0.38	0.39	0.50	0.47	0.46
$T_0 = 50$	0.57	0.55	0.55	0.50	0.47	0.49	0.57	0.57	0.56
$T_0 = 100$	0.61	0.58	0.59	0.56	0.51	0.50	0.61	0.59	0.60

Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$

	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.56	0.55	0.55	0.49	0.50	0.52	0.59	0.58	0.57
$T_0 = 50$	0.59	0.57	0.58	0.54	0.55	0.55	0.62	0.60	0.62
$T_0 = 100$	0.62	0.60	0.60	0.59	0.58	0.55	0.64	0.62	0.63

Power DGP1b (Dense SC Model)

	Synthetic control			Factor model			Constrained Lasso		
i.i.d. data with $\rho_\epsilon = \rho_u = 0$									
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.53	0.53	0.53	0.38	0.43	0.47	0.50	0.50	0.50
$T_0 = 50$	0.59	0.59	0.58	0.52	0.54	0.57	0.58	0.57	0.57
$T_0 = 100$	0.61	0.61	0.60	0.55	0.58	0.60	0.60	0.59	0.60
Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$									
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.63	0.63	0.66	0.52	0.57	0.64	0.61	0.61	0.62
$T_0 = 50$	0.64	0.65	0.67	0.56	0.61	0.62	0.64	0.64	0.66
$T_0 = 100$	0.63	0.65	0.66	0.58	0.60	0.61	0.64	0.64	0.65

Power DGP2a (Factor Model w/ Common Support)

	Synthetic control			Factor model			Constrained Lasso		
i.i.d. data with $\rho_\epsilon = \rho_u = 0$									
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.49	0.52	0.51	0.32	0.41	0.46	0.46	0.48	0.49
$T_0 = 50$	0.55	0.55	0.58	0.45	0.52	0.57	0.55	0.54	0.55
$T_0 = 100$	0.54	0.57	0.59	0.49	0.55	0.60	0.55	0.57	0.60
Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$									
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.60	0.62	0.64	0.49	0.57	0.63	0.58	0.61	0.62
$T_0 = 50$	0.61	0.63	0.65	0.53	0.58	0.62	0.61	0.64	0.65
$T_0 = 100$	0.61	0.63	0.67	0.54	0.59	0.63	0.61	0.63	0.67

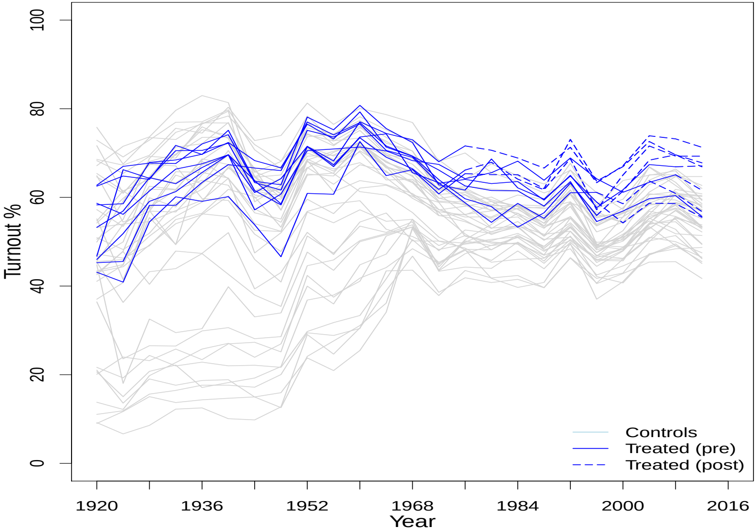
Power DGP2 (Factor Model w/o Common Support)

	i.i.d. data with $\rho_\epsilon = \rho_u = 0$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.16	0.17	0.19	0.21	0.28	0.37	0.37	0.38	0.38
$T_0 = 50$	0.18	0.20	0.21	0.26	0.39	0.46	0.43	0.44	0.45
$T_0 = 100$	0.18	0.22	0.25	0.28	0.40	0.51	0.44	0.47	0.50
	Weakly dependent data with $\rho_\epsilon = \rho_u = 0.6$								
	Synthetic control			Factor model			Constrained Lasso		
	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$	$J = 10$	$J = 20$	$J = 50$
$T_0 = 20$	0.21	0.23	0.28	0.30	0.36	0.45	0.43	0.42	0.46
$T_0 = 50$	0.23	0.25	0.29	0.34	0.43	0.52	0.47	0.48	0.51
$T_0 = 100$	0.22	0.24	0.28	0.35	0.45	0.54	0.47	0.50	0.51

Effect of EDR Laws on Voter Turnout

- Revisit analysis in Xu (2017)
- Background on election day registration (EDR) laws:
    - Voting in US is typically a two step procedure: (1) registration, (2) voting
    - The two steps usually require separate trips $\rightarrow$ costly
    - EDR: allows for registration when arriving at the polling station.
    - Enacted in 1970s: Maine, Minnesota and Wisconsin
    - Enacted in 1990s: Idaho, New Hampshire and Wyoming
    - Enacted before 2012 election: Montana, Iowa and Connecticut
- State-level turnout data, excluding Hawaii, Alaska and North Dakota.
    - From 1920 to 2012, only presidential elections.
- Separate analysis for all nine treated states
- Controls: all states which never introduced EDR

Raw Data

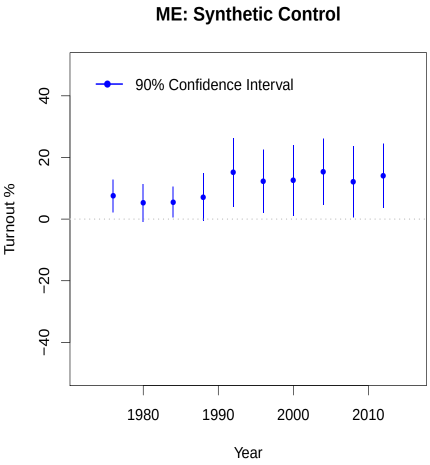

Turnout %
100
80
60
40
20
0
1920
1936
1952
1968
1984
2000
2016
Year
Controls
Treated (pre)
Treated (post)

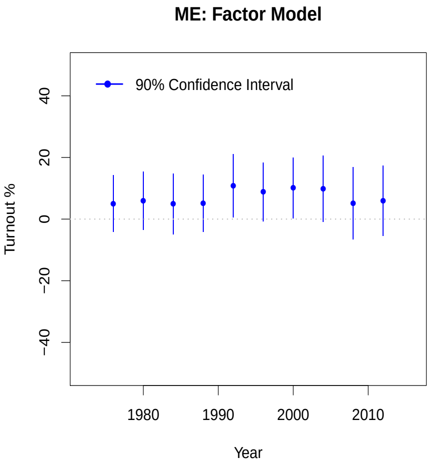
No-Effects Hypothesis

$$H_0 : (\alpha_{T_0+1}, \ldots, \alpha_T) = (0, \ldots, 0)$$

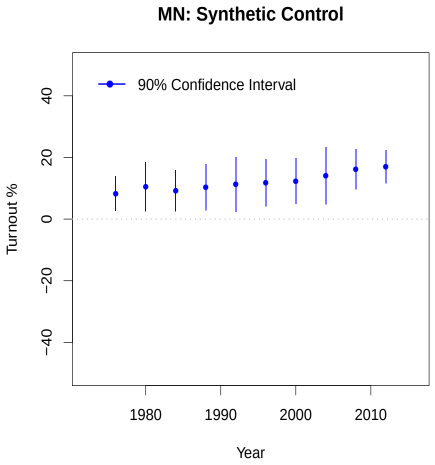
	Moving Block Permutations			i.i.d. Permutations		
	Synth.Control	Factor Model	Constr.Lasso	Synth.Control	Factor Model	Constr.Lasso
CT	0.08	0.29	0.04	0.08	0.29	0.04
IA	0.04	0.25	0.29	0.01	0.20	0.26
ID	0.83	0.04	0.42	0.70	0.04	0.44
ME	0.04	1.00	0.83	0.00	1.00	0.91
MN	0.04	0.96	0.58	0.00	0.93	0.54
MT	0.38	0.33	0.96	0.32	0.26	0.90
NH	0.04	0.21	0.38	0.00	0.09	0.33
WI	0.04	0.92	0.17	0.00	0.72	0.05
WY	0.46	0.25	0.62	0.42	0.37	0.65

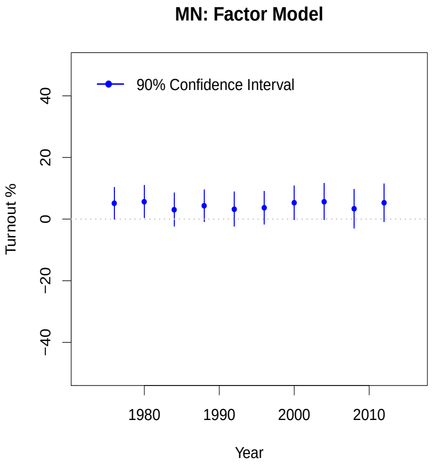
Maine


ME: Synthetic Control
90% Confidence Interval
Turnout %
40
20
0
−20
−40
1980
1990
2000
2010
Year

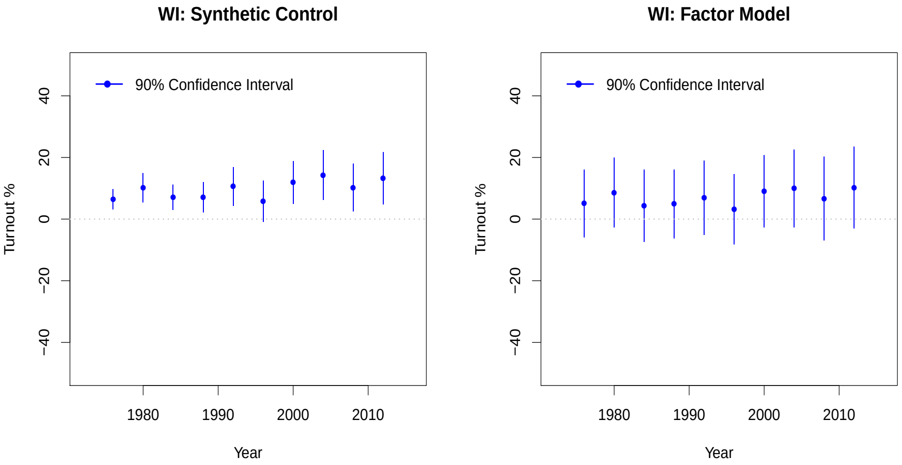

ME: Factor Model
90% Confidence Interval
Turnout %
40
20
0
−20
−40
1980
1990
2000
2010
Year

Minnesota

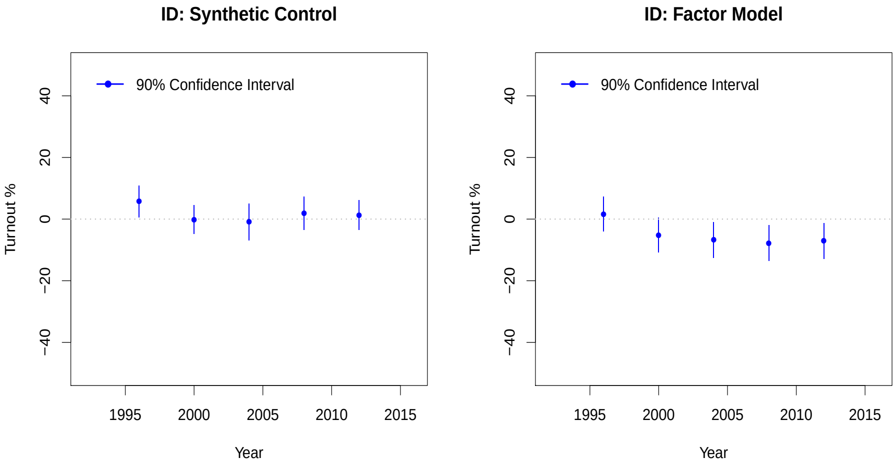

MN: Synthetic Control
90% Confidence Interval
Turnout %
Year
40
20
0
−20
−40
1980
1990
2000
2010


MN: Factor Model
90% Confidence Interval
Turnout %
Year
40
20
0
−20
−40
1980
1990
2000
2010

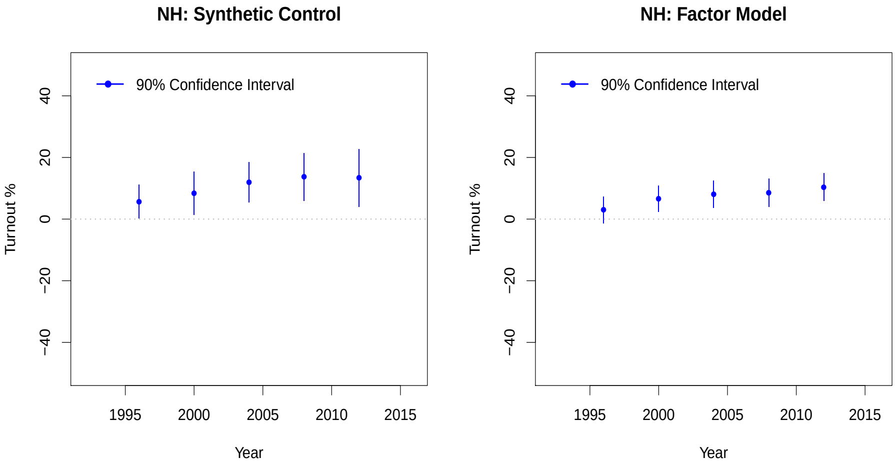
Wisconsin


WI: Synthetic Control
90% Confidence Interval
Turnout %
40
20
0
−20
−40
1980
1990
2000
2010
Year
WI: Factor Model
90% Confidence Interval
Turnout %
40
20
0
−20
−40
1980
1990
2000
2010
Year

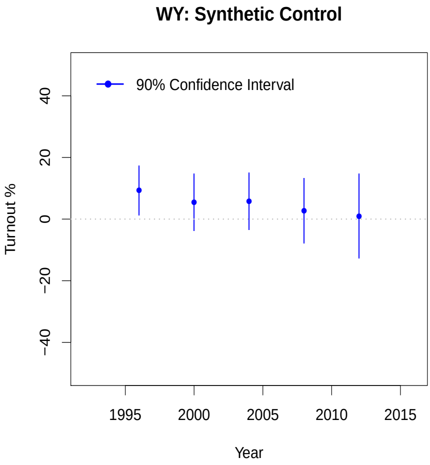
Idaho

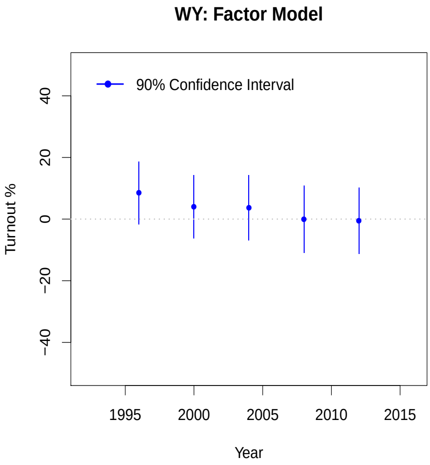

ID: Synthetic Control
90% Confidence Interval
Turnout %
40
20
0
−20
−40
1995
2000
2005
2010
2015
Year
ID: Factor Model
90% Confidence Interval
Turnout %
40
20
0
−20
−40
1995
2000
2005
2010
2015
Year

New Hampshire

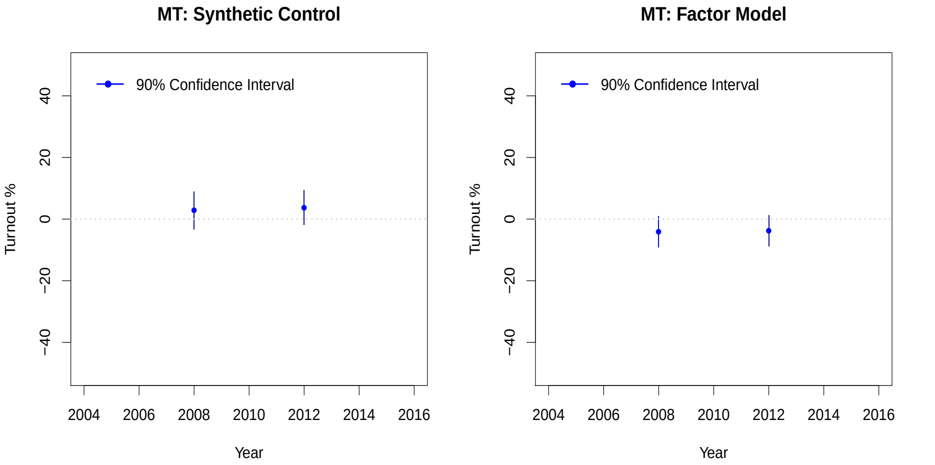


Wyoming

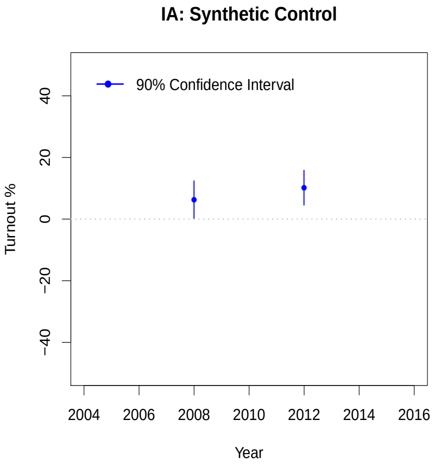

WY: Synthetic Control
90% Confidence Interval
Turnout %
Year
40
20
0
−20
−40
1995
2000
2005
2010
2015


WY: Factor Model
90% Confidence Interval
Turnout %
Year
40
20
0
−20
−40
1995
2000
2005
2010
2015

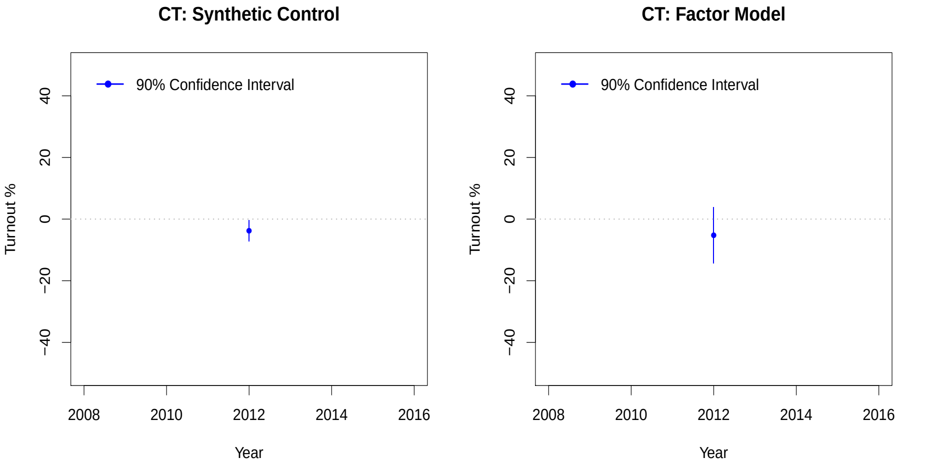
Montana


MT: Synthetic Control
90% Confidence Interval
Turnout %
40
20
0
−20
−40
2004
2006
2008
2010
2012
2014
2016
Year
MT: Factor Model
90% Confidence Interval
Turnout %
40
20
0
−20
−40
2004
2006
2008
2010
2012
2014
2016
Year

Iowa


IA: Synthetic Control
90% Confidence Interval
Turnout %
Year
40
20
0
−20
−40
2004
2006
2008
2010
2012
2014
2016
IA: Factor Model
90% Confidence Interval
Turnout %
Year

Connecticut



Conclusions

- We propose a general inference procedure for counterfactual and synthetic control (CSC) methods.
- The proposed procedure has a double justification.
- The proposed method exploits the times series dimension of the problem.
- The proposed method works in conjunction with many different CSC methods.
- Our procedure works well in finite samples.

Thank you!

ABADIE, A., A. DIAMOND, AND J. HAINMUELLER (2010): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californias Tobacco Control Program," *Journal of the American Statistical Association*, 105(490), 493–505.

——— (2015): "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59(2), 495–510.

ABADIE, A., AND J. GARDEAZABAL (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country," *The American Economic Review*, 93(1), 113–132.

ANDREWS, D. W. (2003): "End-of-Sample Instability Tests," *Econometrica*, 71(6), 1661–1694.

ATHEY, S., M. BAYATI, N. DOUDCHENKO, G. IMBENS, AND K. KHOSRAVI (2017): "Matrix Completion Methods for Causal Panel Data Models," .

CARVALHO, C. V., R. MASINI, AND M. C. MEDEIROS (2017): "Arco: an artificial counterfactual approach for high-dimensional panel time-series data," .

CHAN, M., AND S. KWOK (2016): "Policy Evaluation with Interactive Fixed Effects," .

DOUDCHENKO, N., AND G. W. IMBENS (2016): "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," Working Paper 22791, National Bureau of Economic Research.

FIRPO, S., AND V. POSSEBOM (2017): "Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets," .

GOBILLON, L., AND T. MAGNAC (2016): "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls," *The Review of Economics and Statistics*, 98(3), 535–551.

HAHN, J., AND R. SHI (2016): "Synthetic Control and Inference," Mimeo.

HSIAO, C., H. STEVE CHING, AND S. KI WAN (2012): "A PANEL DATA APPROACH FOR PROGRAM EVALUATION: MEASURING THE BENEFITS OF POLITICAL AND ECONOMIC INTEGRATION OF HONG KONG WITH MAINLAND CHINA," *Journal of Applied Econometrics*, 27(5), 705–740.

LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing statistical hypotheses*. Springer Science & Business Media.

LEI, J., M. G'SELL, A. RINALDO, R. J. TIBSHIRANI, AND L. WASSERMAN (2017): "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, (just-accepted).

LI, K. (2018): "Inference for Factor Model Based Average Treatment Effects," .

LI, K. T. (2017): "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods," .

ROMANO, J. P. (1990): "On the Behavior of Randomization Tests Without a Group Invariance Assumption," *Journal of the American Statistical Association*, 85(411), 686–692.

VOVK, V., A. GAMMERMAN, AND G. SHAFER (2005): *Algorithmic Learning in a Random World*. Springer.

VOVK, V., I. NOURETDINOV, AND A. GAMMERMAN (2009): "On-line predictive linear regression," *The Annals of Statistics*, 37(3), 1566–1590.

XU, Y. (2017): "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 25(1), 57–76.